Measuring the Sensitivity of Single-Locus "Neutrality Tests" Using a Direct Perturbation Approach

Measuring the Sensitivity of Single-locus "Neutrality Tests" Using a Direct Perturbation Approach

Daniel Garrigan,*† Richard Lewontin, and John Wakeley

Department of Organismic and Evolutionary Biology, Harvard University

***Corresponding author:** E-mail: daniel.garrigan@rochester.edu.

†Present address: Department of Biology, University of Rochester.

Associate editor: John H. McDonald

Abstract

A large number of statistical tests have been proposed to detect natural selection based on a sample of variation at a single genetic locus. These tests measure the deviation of the allelic frequency distribution observed within populations from the distribution expected under a set of assumptions that includes both neutral evolution and equilibrium population demography. The present study considers a new way to assess the statistical properties of these tests of selection, by their behavior in response to direct perturbations of the steady-state allelic frequency distribution, unconstrained by any particular nonequilibrium demographic scenario. Results from Monte Carlo computer simulations indicate that most tests of selection are more sensitive to perturbations of the allele frequency distribution that increase the variance in allele frequencies than to perturbations that decrease the variance. Simulations also demonstrate that it requires, on average, $4N$ generations (N is the diploid effective population size) for tests of selection to relax to their theoretical, steady-state distributions following different perturbations of the allele frequency distribution to its extremes. This relatively long relaxation time highlights the fact that these tests are not robust to violations of the other assumptions of the null model besides neutrality. Lastly, genetic variation arising under an example of a regularly cycling demographic scenario is simulated. Tests of selection performed on this last set of simulated data confirm the confounding nature of these tests for the inference of natural selection, under a demographic scenario that likely holds for many species. The utility of using empirical, genomic distributions of test statistics, instead of the theoretical steady-state distribution, is discussed as an alternative for improving the statistical inference of natural selection.

Key words: DNA sequence, infinite-allele model, infinite-sites model, natural selection, polymorphism.

Research article

Introduction

A major question pertinent to understanding the genetic variation within and between species is how important natural selection has been in determining that variation. That is, how much of the observed genetic differentiation is a consequence of direct physiological, developmental, and behavioral causal relations between DNA sequence variation and variation in fertility and probability of survival of individuals carrying these sequences. The obvious direct approach to this problem would be to measure the components of reproductive fitness in different genotypes, but there are a number of serious limitations inherent in this approach, especially for animals (Orr 2009). First, it is extremely difficult, if not impossible, to obtain complete fitness components, including probabilities of different matings, fertility schedules, and life tables for organisms whose life cycle cannot be observed in detail under natural conditions. Second, even if complete life cycle components can be obtained, large enough sample sizes to detect selection are not possible unless fitness effects are drastic (e.g., see the study by Christiansen and Frydenberg [1973] of an esterase polymorphism in the live-bearing fish *Zoarces viviparus*, where all the components of fitness could be measured). Third, measures of fitness components in present-day environments may not necessarily reflect fitness in past environments.

One solution offered to circumvent these difficulties has been to attempt to infer the occurrence of natural selection from extant patterns of standing genetic variation in the genome of one or more populations. This has taken the form of a large number of statistical tests based on the "null hypothesis" that there has been no natural selection and that the frequency distribution of variants in a sample can be predicted under the assumptions that the population is at a stochastic steady state expected in populations of a fixed breeding size and fixed model of mutation. If a sample shows a statistically significant deviation from the expected theoretical distribution, the null hypothesis that there is no selection is rejected.

The problem with this indirect approach to detect selection is one shared by all statistical procedures that claim to test a null hypothesis. What these statistical procedures actually test is not a null hypothesis, but rather, a null "structure"—a set of claims about a universe, one claim being isolated as the "hypothesis" and remainder being relegated to the category of "assumptions." The test procedures themselves do not distinguish between the hypothesis and the assumptions. For example, until computer simulations became possible, it was not known

whether the *t*-test for the difference between means was very sensitive to the assumption of normality or whether the equality of variances of the distributions was critical to the test. As it turned out, the test is robust against both non-normality and unequal variances. However, the *F* test for the equality of variances from two populations turns out to be extremely sensitive to the assumption of underlying normal distributions of the variable. To validate various tests for "selection," a similar program of examining their sensitivity to the assumptions needs to be carried out systematically. The question becomes how such a systematic examination is to be carried out.

The usual set of assumptions inherent in the null structure of neutral evolution includes:

1) a constant mutation rate in which each mutation is to a new allele (infinite-alleles tests) or at a new site in DNA sequence (infinite-sites tests);
2) a constant breeding size;
3) no effective migration into the population from populations that are not, themselves, at stochastic steady state;
4) a sufficient elapsed time since the foundation of the population, with the current values of mutation rate and population size, so that a stochastic steady state has been reached; and
5) assumptions about linkage between sites (either total or no linkage disequilibrium between sites).

It has been widely understood for some time that various tests for selection may give misleading results because at least one of the assumptions has not been fulfilled (Ewens and Gillespie 1974; Tajima 1989a; Sawyer and Hartl 1992; Fu 1997; Kreitman 2000; Nielsen 2001; Ramos-Onsins and Rozas 2002; Eyre-Walker 2002; Nielsen 2005; Thornton 2005; Sjödin et al. 2005; Kim 2006; Jensen et al. 2007; Zeng, Shi, and Wu 2007; Ramírez-Soriano et al. 2008). It is of particular interest that the paper of Tajima (1989a), which points out the problems that nonequilibrium demography pose to "tests of selection," appeared immediately after the original paper proposing the measure *D* (Tajima 1989b) in the same issue of *Genetics*. It is not the intent of the present paper to review and analyze this vast literature in detail. All these studies use essentially the same methodology. Realizing that demographic events may affect the pattern of nucleotide variation, they assess the sensitivity of a test for selection to deviations in the form of particular demographic scenarios. For example, in their seminal study, Simonsen et al. (1995) simulated population size bottlenecks or sudden expansions or subdivision of an original population into 2 subpopulations. The results of this and other studies show that various tests are, indeed, sensitive to demographic changes and may give significant results in the absence of selection.

It is already known that, for specific tests of selection, it is possible to find demographic events (such as population bottlenecks) that will produce statistically significant test results for a specific test procedure. However, the fact that one can find such a specific scenario for a specific test tells us virtually nothing about the general usefulness of the test procedures. For example, it has been well known for 50 years that the conventional *t*-test for the difference between means is generally robust to the assumption of normality of the underlying distributions, but one can find very unusual distributions that give badly biased results. To obtain a picture of the general reliability of the great variety of tests of selection, it would seem necessary to apply an immense of array of demographic scenarios for each test.

The problem with this approach to determining the robustness of tests for selection is the immense size of the model and parameter space that needs to examined. In the case of changes in population size, this space includes not only the time since the given expansion or contraction, and the size of the population before and after an event, but the historical sequence of repeated events, each with different size and duration parameters. Feasible approaches to explore this space are usually limited to considering a small number of nonequilibrium demographic events in the context of an otherwise "steady-state" demographic model. On the other hand, stochastic demographic models can converge to equilibrium predictions under the condition that the rate of demographic fluctuation is fast relative to the rate of the ancestral process (Sjödin et al. 2005). The dimensionality of the underlying problem makes any reasonable examination of the parameter space impractical, especially, in view of the known sensitivity of various tests to at least a few examples of such perturbations. What is needed for general assessment of the test procedures is an approach that is not bound to a particular example but rather provides a more complete picture of the sensitivity of the tests to a range of nonselective perturbations.

An alternative approach is to set aside particular biological scenarios and, instead, to examine the sensitivity of the tests to direct perturbations of varying amounts from the steady-state frequency distributions derived from the full null structure. In the case of a dynamical process, such an examination includes both the effect of perturbations of various degrees in the shape of the distribution and the relaxation behavior of the statistical tests as the perturbed distributions relax toward their stochastic equilibrium. This approach, which became standard in statistics after the introduction of Monte Carlo computer modeling made the investigation possible, separates what is actually being done in a statistical test, which is to test whether observations are a good fit to some probability distribution, from questions of which elements in the null structure that generated the distribution are responsible for the deviation. Here, this sensitivity analysis approach is developed for tests of selection that use genetic variation at a single locus.

Materials and Methods

Simulation Methodology and Tests of Selection

The investigation was carried out by Monte Carlo computer simulation of populations and samples. Because computer simulation results can be critically sensitive to small, undetected programming errors or hidden differences in assumptions, population simulations were carried

out independently by each of the authors, using different programming languages on different computer systems. In one case, a counterintuitive result was first found in the infinite-alleles relaxation tests, but this phenomenon was then verified in independently programmed infinite-sites models. Most cases were simulated by forward simulation of population generations and some by backward coalescent simulation. It is also important to note that none of the simulation protocols incorporate the effects of intralocus recombination.

A sampling of standard tests for selection that have been designed for protein or DNA sequence data were then applied to samples of various sizes from the simulated populations. These tests all utilize the frequency distribution of mutations as a common foundation: for infinite-alleles models, the Ewens–Watterson *F* test (Watterson 1978); for infinite-sites models, Fu and Li's *D* and *F* tests (Fu and Li 1993), Tajima's *D* test (Tajima 1989b), Fu's F_S test (Fu 1996), and Fay and Wu's *H* test (Fay and Wu 2000). For infinite-alleles test simulations, every nucleotide sequence copy in the population was initially given the serial number 1, and then, as mutations accumulated, the allele copy on which the most recent mutation occurred was reidentified by the serial number of this most recent mutation. For infinite-sites test simulations, sequences were modeled as an *n*-ary tree structure, in which nodes contained the count of the sequence in the population and nodes that went extinct were removed from the data structure only if they had no descendant sequences in the population.

Because the methodology of the study involves the results of tests on samples from populations generated by Monte Carlo simulation with various perturbations from the null structure, test stochastic steady-state null populations were generated by Monte Carlo simulation as a control for the simulation programs. The rule used in the main body of our study for generating steady-state populations for different population sizes and mutation rates was to begin with a single-individual gene sequence and run the simulations until the mutation and random drift process had eliminated the last copy of the original sequence. For an infinite-alleles model, the expected number of generations until a population beginning with a single allele will be essentially at steady state is given by Ewens and Gillespie (1974) as the number of generations, *t*, such that the *t*th power of leading eigenvalue of the stochastic process

$$\lambda_1^t = [1 - \frac{(\theta + 1)}{2N}]^t = 0.01, \text{yielding}, \quad (1)$$

$$t = \frac{9.2N}{(\theta + 1)}, \quad (2)$$

in which $\theta = 4N\mu$.

In our 100 control simulations of the parameters $2N = 5{,}000$, $\mu = 2 \times 10^{-4}$, and $2N = 10^4$, $\mu = 10^{-4}$, the observed mean number of generations to simulated steady state and their standard errors (SEs) were 10,731 ± 558.9 and 20,658 ± 1,238.8, respectively. These means are much larger than the expected values of 7,666 and 15,333, but the large SEs allow for the possibility that an occasional simulated population will not meet the very stringent criterion given in equation (2). A second control on the simulations compares the numbers of alleles (*k*) observed in a sample of size *n* to the expected number of alleles sample from a steady-state population from the relation

$$E(k) = \theta \sum_{i=0}^{n-1} \frac{1}{(\theta + i)}. \quad (3)$$

(Ewens 1969).

The major control on the simulation implementation, in general, was to draw between 2,000 and 6,000 small samples, each drawn from a separately simulated replicate steady-state null population. Each batch of replicated populations, and the samples from them, had a different combination of values for the effective number of chromosomes ($2N$), the mutation rate (μ), and sample size (n). Each sample then provided data for the statistical test procedures proposed by Ewens (1972) for the infinite-alleles model, for which probabilities were determined by Monte Carlo simulation. The objective of these control simulations was to compare the observed distribution of test probabilities with the expected null distribution. In the case of the infinite-sites tests, the above objective was achieved using *P* values generated via Monte Carlo simulation of the coalescent process, with population mutation rates estimated from the simulated data. For each simulated data set, the *P* value under a standard neutral model was estimated from 5,000 coalescent replicates.

Three Phases of the Study

The study of the behavior of the various statistical tests under deviations from the null structure was carried out in three phases:

1) perturbations from the null structure frequency distribution;
2) relaxation times from initial frequency conditions; and
3) an example of a common demographic cycle.

The first two phases, unlike the usual examination of the sensitivity of tests for selection to particular demographic events, are not in terms of specific historical models of population size variation but deal only with deviations of the allelic frequency distribution from the theoretical long-term steady state. Only in the third phase is a common biological scenario introduced to illustrate a case of cyclical demographic and migration structure.

In the first phase, perturbations of various amounts were made from the theoretical steady-state distribution of allele frequencies that arise from the null assumptions of no selection, constant population size, fixed mutation rate, and no migration into the population from other populations with different allelic frequency distributions. The steady-state distribution of allelic frequencies (x) under the assumption that every mutation leads to a different state

is given by

$$\phi(x) = C\theta(1 - x)^{\theta - 1}x^{-1}, \quad (4)$$

in which C is a normalizing constant (Kimura and Crow 1964). In equation (4), when $\theta \geq 1$, the allelic frequency distribution is expected to be L shaped, predicting one allele with high frequency and a number of alleles with low or very low frequency. The resulting allelic distributions were altered to make them more or less asymmetrical, whereas remaining L shaped, by a simple rule of allele frequency perturbation.

In a typical example from our simulations, with $2N = 5{,}000$ and $\mu = 2 \times 10^{-4}$, there are 14 alleles with one allele frequency = 0.58, five alleles between 0.1 and 0.01, and eight alleles with frequencies between 0.01 and 10^{-4}. The perturbation method consisted in moving the frequency of the highest frequency allele or site mutation toward or away from fixation by a fraction, q, of its distance from 1.00 and then readjusting the frequencies of the remaining alleles proportionately so that the frequencies sum to unity. Thus, with n unperturbed allele frequencies (x_i), ordered from smallest to largest, the allele frequencies after perturbing with intensity P are

$$x_i' = x_i(1 - P), \quad (5)$$

for $i = 1$ to $n - 1$ and

$$x_n' = P(1 - x_n), \quad (6)$$

where P is positive for "positive" perturbations that move the highest unperturbed frequency closer to fixation at unity and the remainder of the frequencies closer to zero, whereas "negative" perturbations move the highest unperturbed frequency downward and the remaining allele frequencies upward.

As a simplified numerical example, the frequencies {0.7, 0.2, 0.1}, if positively perturbed with intensity +0.6, would be changed to {0.88, 0.08, 0.04}, whereas a negative perturbation of −0.6 would result in the array {0.52, 0.32, 0.16}. The consequence is that positive perturbations make the L-shaped allele frequency distribution more pronouncedly L-shaped with a higher variance of allele frequency, whereas negative perturbations move all the allele frequencies closer together with a consequent lower variance among the frequencies. This form of perturbation is meant to change the shape of the allele frequency distribution only moderately without, for example, changing the original ordering of the frequencies. Given the observed L-shaped distributions generated by our implementation of the null model, perturbations within the range +0.6 to −0.6 had this effect. These upper and lower bounds for the perturbation coefficients were determined to sufficiently skew the allelic frequency distribution, whereas more extreme positive values had the effect of eliminating all polymorphism from the population sample. Samples were then taken from the perturbed population frequencies, and statistical tests were then performed to determine the sensitivity of the tests to such perturbations. In the case of the infinite-sites tests, the sequence identity of the alleles, whose frequencies were to be perturbed, were initially configured by randomly generating gene trees under a standard neutral model.

Table 1. Observed Mean and Expected Number of Alleles at Estimated Steady State from Control Simulations, with Haploid Population Size 2*N*, Per-Locus Mutation Rate μ, and Per-Locus Population Mutation Rate θ.

2N	μ	θ	Observed Mean	Expected
5,000	0.002	20	114.6	111.0
10,000	0.0002	4	36.6	31.8
5,000	0.0002	2	16.2	16.2
10,000	0.0001	2	17.7	17.6
5,000	0.0001	1	8.8	8.5

In the second phase of the study, populations of fixed size and mutation rate were initiated from a variety of genetic compositions that were conceived to be very different from the eventual steady state. The process of relaxation toward the steady state was followed by performing the statistical tests for the null structure at various times during the relaxation process. In particular, we were interested in the time needed before the trace of the original strong perturbation was lost in the statistical tests. For any dynamical process that tends to an equilibrium state from other points in the state space, a critical issue is the speed of the approach to the equilibrium from incidents of perturbation. If the rate of relaxation toward equilibrium is fast relative to the temporal frequency of perturbations, then, even if those perturbations are large, it is to be expected that the state of the system will be very close to equilibrium when assessed at random times. On the other hand, slow relaxation rates mean that the system state is likely to carry the marks of even old perturbations. The relaxation time problem is crucial for our understanding of the results of tests for selection because we would like to distinguish the deviations from the stochastic steady state that arise from selection from those that result from a history of changes in demographic and migration events.

We have studied the relaxation problem by simulating populations that begin with allele frequency distributions that have a variety of shapes more or less different from the steady-state distribution and following the relaxation from these initial distributions for various population sizes, mutation rates, and sample sizes. The initial allele frequency distributions were not chosen to reflect some particular demographic histories but were meant to span a diversity of deviant starting conditions in order to determine how dependent the relaxation times are on starting in different regions of the state space of the initial non–steady-state distribution. The initial distributions examined were monomorphism, two equally frequent alleles and, 10 alleles in the roughly exponentially decreasing L-shaped distribution, with a vector of allele frequencies equal to {0.5, 0.24, 0.12, 0.06, 0.03, 0.01, 0.01, 0.01, 0.01, 0.01}. Haploid population sizes were 2,000, 5,000, and 10^4, and sample sizes were 50 and 100, with $\theta = 2$ and a few cases with $\theta = 6$. For the infinite-sites model, initial sequence

Table 2. Verification Runs for Samples from Monte Carlo Simulated Populations (*n* is the sample size).

					Proportion of Tests Falling at Different Theoretical Probability Levels							
					Lower Tail				Upper Tail			
θ	2N	μ	*n*	Replicates	0.001	0.010	0.025	0.050	0.050	0.025	0.010	0.001
2.0	5,000	0.002	100	6,000	0.0020	0.0148	0.0302	0.0595	0.0679	0.0342	0.0145	0.0018
	2,000	0.0005	50	5,000	0.0014	0.0124	0.0258	0.0554	0.0416	0.0180	0.0074	0.0004
			100	3,000	0.0017	0.0093	0.0273	0.0560	0.0485	0.0293	0.0130	0.0030
	1,000	0.001	100	5,000	0.0020	0.0112	0.0312	0.0596	0.0257	0.0396	0.0186	0.0022
1.0	2,000	0.00025	100	5,000	0.0018	0.0161	0.0326	0.0587	0.0576	0.0228	0.0088	0.0016
	1,000	0.0005	50	2,000	0.0020	0.0117	0.0271	0.0516	0.0322	0.0184	0.0061	0
0.5	2,000	0.000125	100	5,000	0.0018	0.0160	0.0340	0.0582	0.0299	0.0075	0.0025	0.0007
	1,000	0.00025	100	5,000	0.0029	0.0164	0.0352	0.0660	0.0321	0.0104	0.0060	0.0020

configurations for the two equally frequent allele condition was determined by randomly generating two sequences that differed by a Poisson-distributed number of mutations with mean θ.

In the final phase of the study, we examine a particular cyclical biological scenario that illustrates how a common form of the regular life history of populations affects the results of the tests in the absence of any selection.

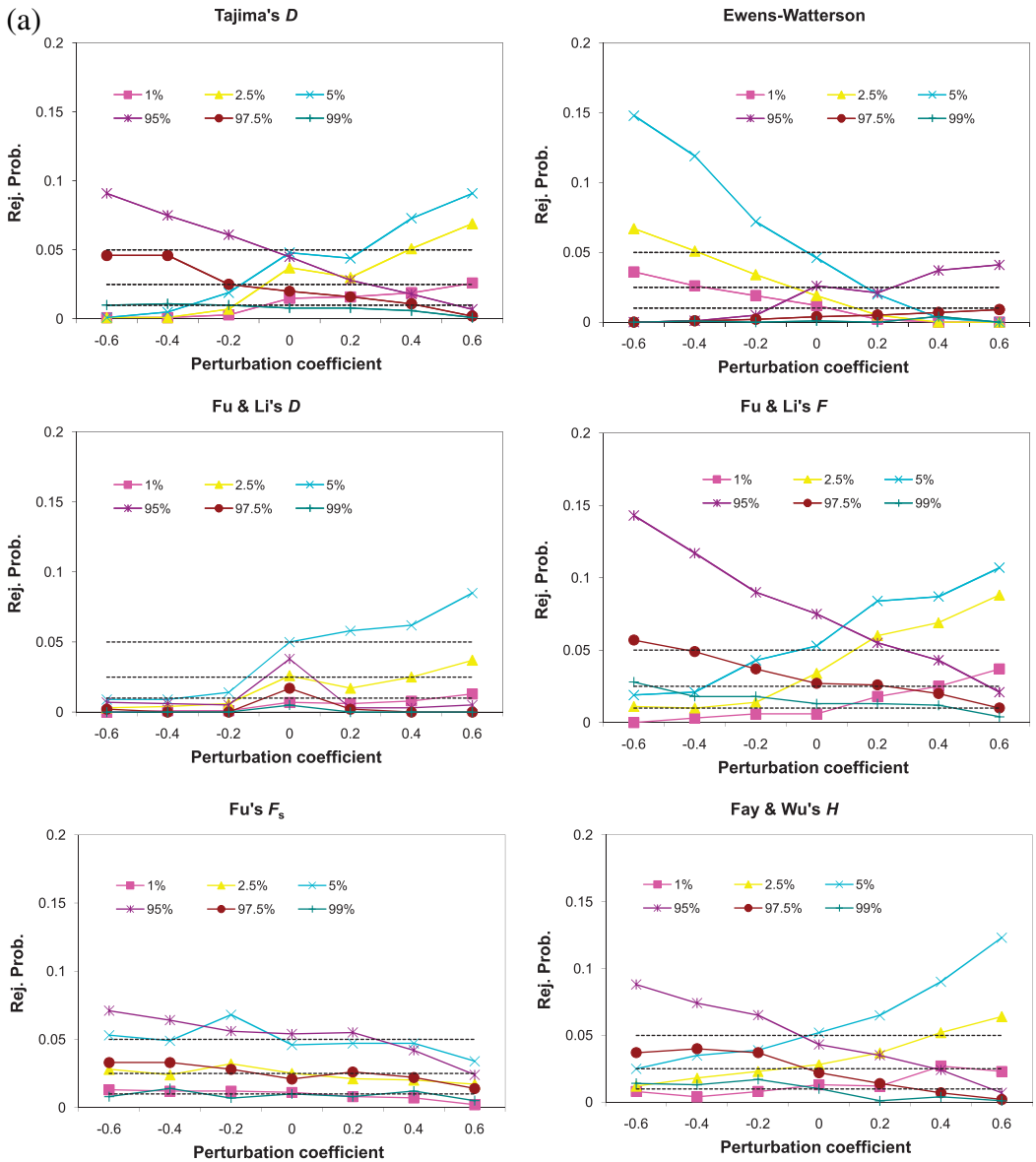

Fig. 1. Observed rejection probabilities for the six tests for different amounts of perturbation from the null distribution. Abscissa: perturbation intensity; ordinate: observed rejection probability; dashed horizontal lines nominal rejection levels. The sample size is $n = 20$; (a) $\theta = 1$, (b) $\theta = 5$, and (c) $\theta = 10$.

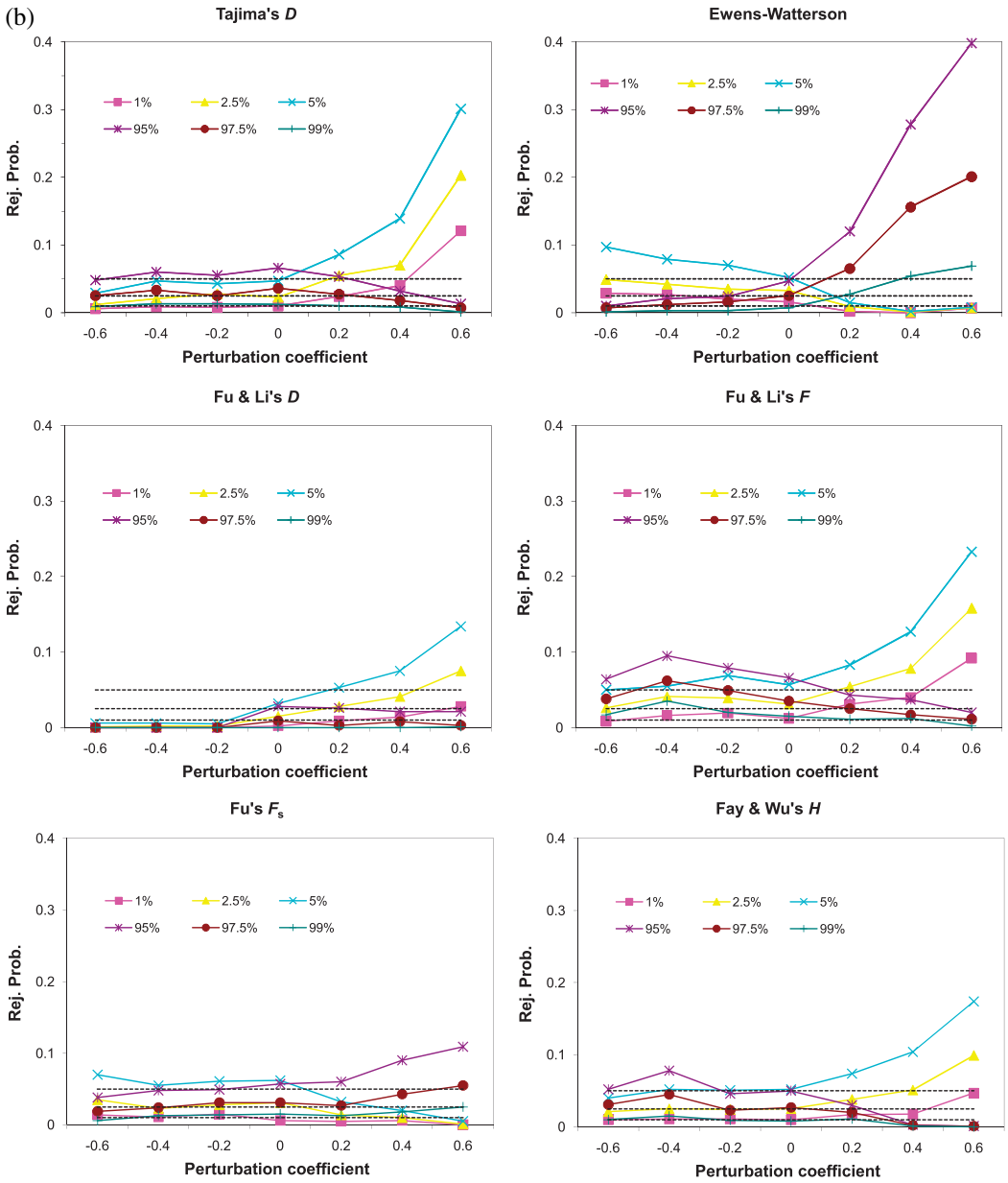

FIG. 1. (Continued).

Results

The robustness of statistical tests of selection to deviations in the null structure was examined in three phases. However, we first report on control simulations that were performed to assess the reliability of the computer programs for recovering the steady-state probability distributions for the various tests of selection. The comparison of the simulated samples with the expected number of alleles (table 1) shows excellent agreement. Furthermore, table 2 compares the simulated probability distribution with the predicted theoretical values. The results show generally good agreement with the expectations, but in almost all cases, there is a small excess of observed test values in the tails of the distributions.

Perturbation Tests

For all the infinite-sites tests and the infinite-alleles test, perturbations of magnitude $-0.6 \leq q \leq +0.6$ were carried out on populations, using the following sets of parameters: $2N = \{2{,}000, 5{,}000\}$, $\theta = \{1, 5, 10\}$, and $n = \{20, 40\}$. Illustrative examples of the sensitivity of the tests to perturbations are given in figures 1 and 2, with $\theta = 1$, 2, and 10 and $2N = 5{,}000$. In general, there is an excess of rejection probability beyond the nominal test probabilities when $q \geq 0.4$, with the true probability of rejection being as much as three to four times higher. With larger values of θ, the sensitivities to negative perturbations disappear, but the effect when there is positive perturbation becomes stronger. The Ewens–Watterson F test is particularly sensitive to these positive perturbations, having rejection probabilities of

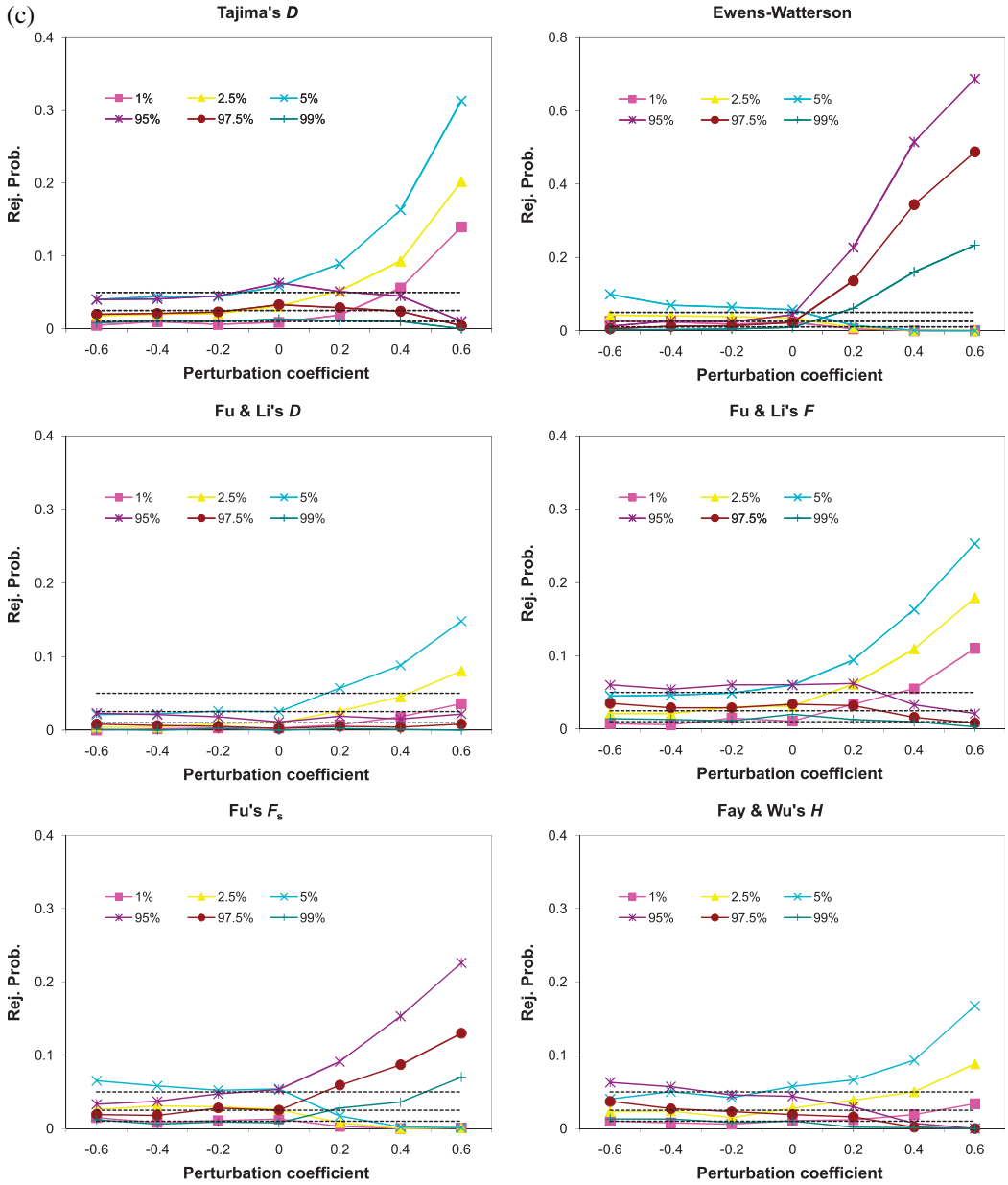

FIG. 1. (Continued).

7–15 times the nominal significance levels for large positive perturbations. Fu's F_S test, on the other hand, appears to be relatively insensitive to all such perturbations. The only test to consistently show sensitivity to negative perturbation is Fu and Li's D, which uses the number of singleton mutations as the basis of the statistic. For all values of θ, but especially for large θ, the tests are more likely to detect an excessively high frequency of the most common allele and a reduction in the low allele frequencies, compared with the null structure. It is this pattern that must then be explained in terms of the possible forces of selection and of demographic and migration histories.

Relaxation Times

Twenty-seven cases involving various combinations of parameter values were examined for relaxation time to the steady state, as determined by the various test statistics. We report the details of some examples and the trends seen as one or another parameter was altered. The examples reported in detail show the range of effects and phenomena observed. What is shown in the following figures and tables are the observed proportions of samples that fell in various percentiles of the test statistic distributions at various times after the perturbation, as compared with the theoretical probabilities under the null structure.

Figures 3 and 4 show the overall relaxation of the test statistic distributions for $2N = 5{,}000$, $\theta = 2$, $n = 50$, for the cases of relaxation from the two initial conditions of monomorphism or two equally frequent alleles. In general, as demonstrated in the figures, it requires on the order of $2N$ generations for the perturbed populations to relax reasonably close to the steady-state predictions. In the case of

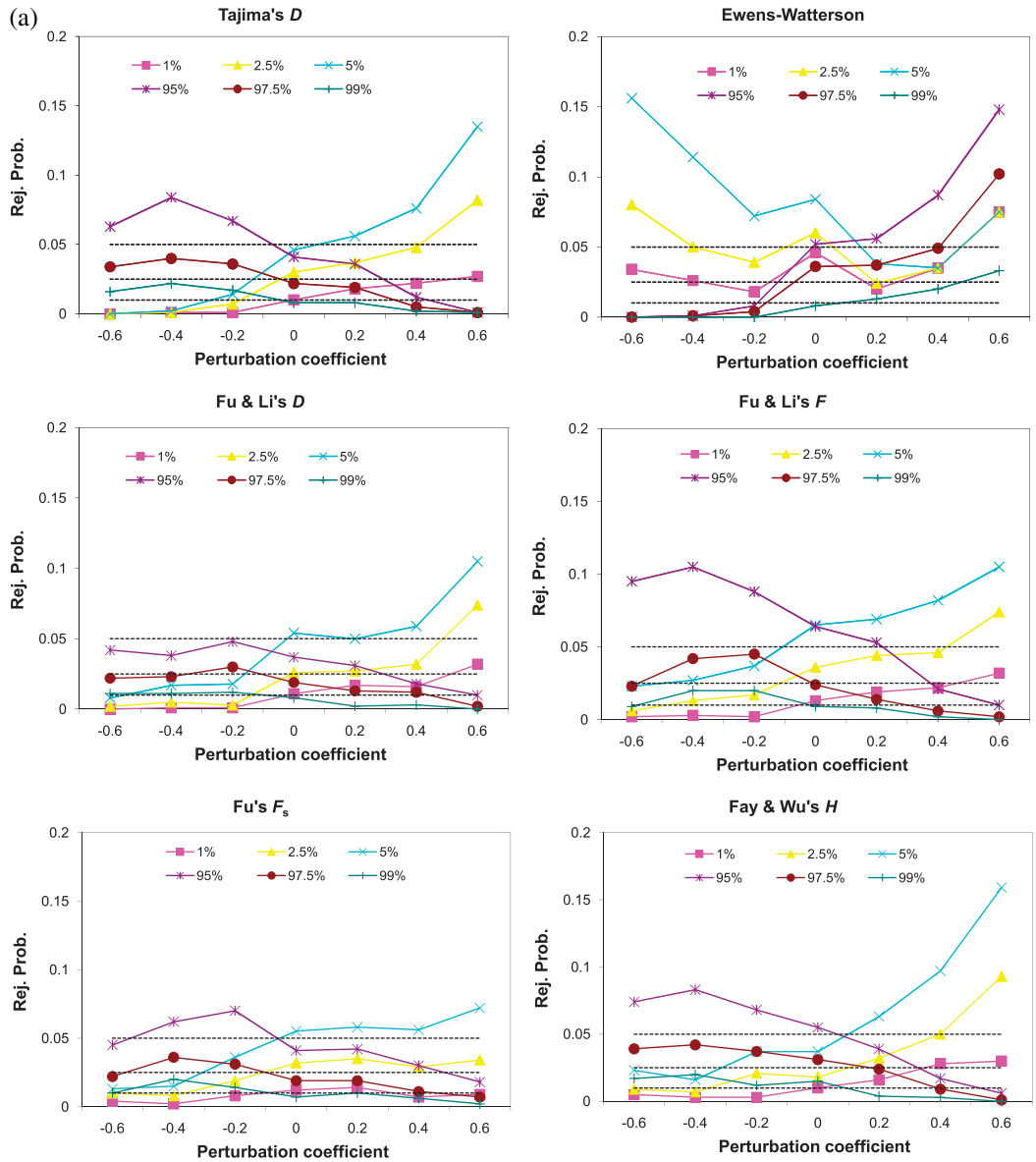

FIG. 2. Observed rejection probabilities for the six tests for different amounts of perturbation from the null distribution. Abscissa: perturbation intensity; ordinate: observed rejection probability; dashed horizontal lines nominal rejection levels. The sample size is $n = 40$; (a) $\theta = 1$, (b) $\theta = 5$, and (c) $\theta = 10$.

the initial L-shaped distribution, which is close in form to the steady-state distribution, there remains a deviation from expected values of about 20% of expectation after 4,000 generations of relaxation (results not shown). As the figures show, the deviations from expected probabilities are in the opposite directions in the case of an initial excessively asymmetrical allele frequency distribution (near monomorphism) as compared with excessively symmetrical perturbation (two equally frequent alleles). The other feature of the general relaxation is that the deviations from the expected test probabilities do not always lie on the same side of the theoretical cumulative probability line. For example, in the case of two initially frequent alleles, Fay and Wu's *H* has excess cumulative probabilities for the first 7,000 generations, but the cumulative frequency curve then passes across the line of equality between observed and expected frequencies and shows a deficiency of cumulative frequencies as it approaches its final stochastic steady state (fig. 4).

The real issue for inferences about selection is not how the overall distributions of the test statistics deviate from the theoretical values during the relaxation process, but what the behavior of the test statistics are in the upper and lower tails of these distributions where statistical significance is judged. Tables 3 and 4 show the results in detail at the 0.05, 0.025, and 0.01 probability levels in the upper and lower tails of the distributions in figures 3 and 4. The data in the tables are given for generation intervals until the asymptotic relaxation process gives observed probabilities of falling in the intervals no greater than about 1.5 times the theoretical values. The tables show that generally it requires between 2*N* and 4*N* generations of relaxation from

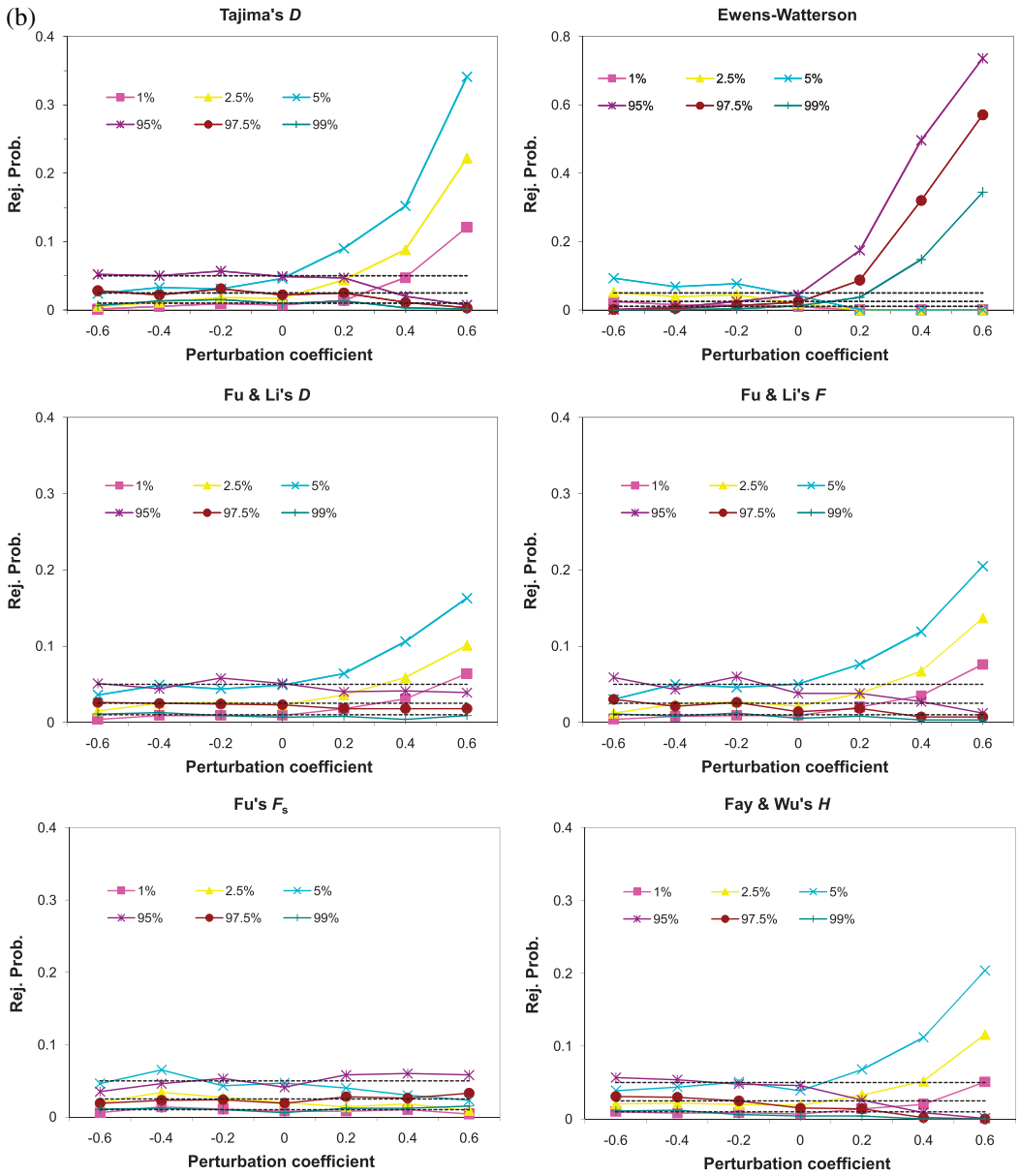

FIG. 2. (Continued).

perturbation before the tests do not give significant deviations from expectation. For example, after 3,000 generations of relaxation from monomorphism, Tajima's *D* is significant 14% of the time at the upper 0.05 significance level, whereas Fu's F_S is significant 6% of the time at the 0.01 significance level after 4,000 generations of relaxation from the two-allele state (table 3). The two tables demonstrate that these results are general over the course of thousands of generations. Exceptions are the Ewens–Watterson *F* test and Fay and Wu's *H*, which show no power to detect deviations during the relaxation from monomorphism, but are sensitive to the relaxation process from the two-allele state at the upper 0.05 and 0.025 levels for between 4,000 and 7,000 generations. The general result is that relaxation from perturbations from the stochastic steady state is a slow process, so that the initial perturbation is still detectable after roughly 2*N* generations.

Cyclical Demographic Patterns—The Ives Scenario as an Example

The stated plan of this investigation was to determine the sensitivity of tests for selection to other deviations from the null structure by making a variety of alterations of the stochastic steady distribution of allele frequencies without specifying particular events that caused those deviations. The results, thus far, suggest that species whose regular demographic pattern includes alternation of population size or repeated immigrations from other populations are likely to deviate significantly from the null structure. For example, a large fraction of species, especially those who reproduce more than once a year and live in a temperate zone with significant cycles of temperature, moisture, and food availability might undergo a large cyclic change in population size and migration rate on a regular basis.

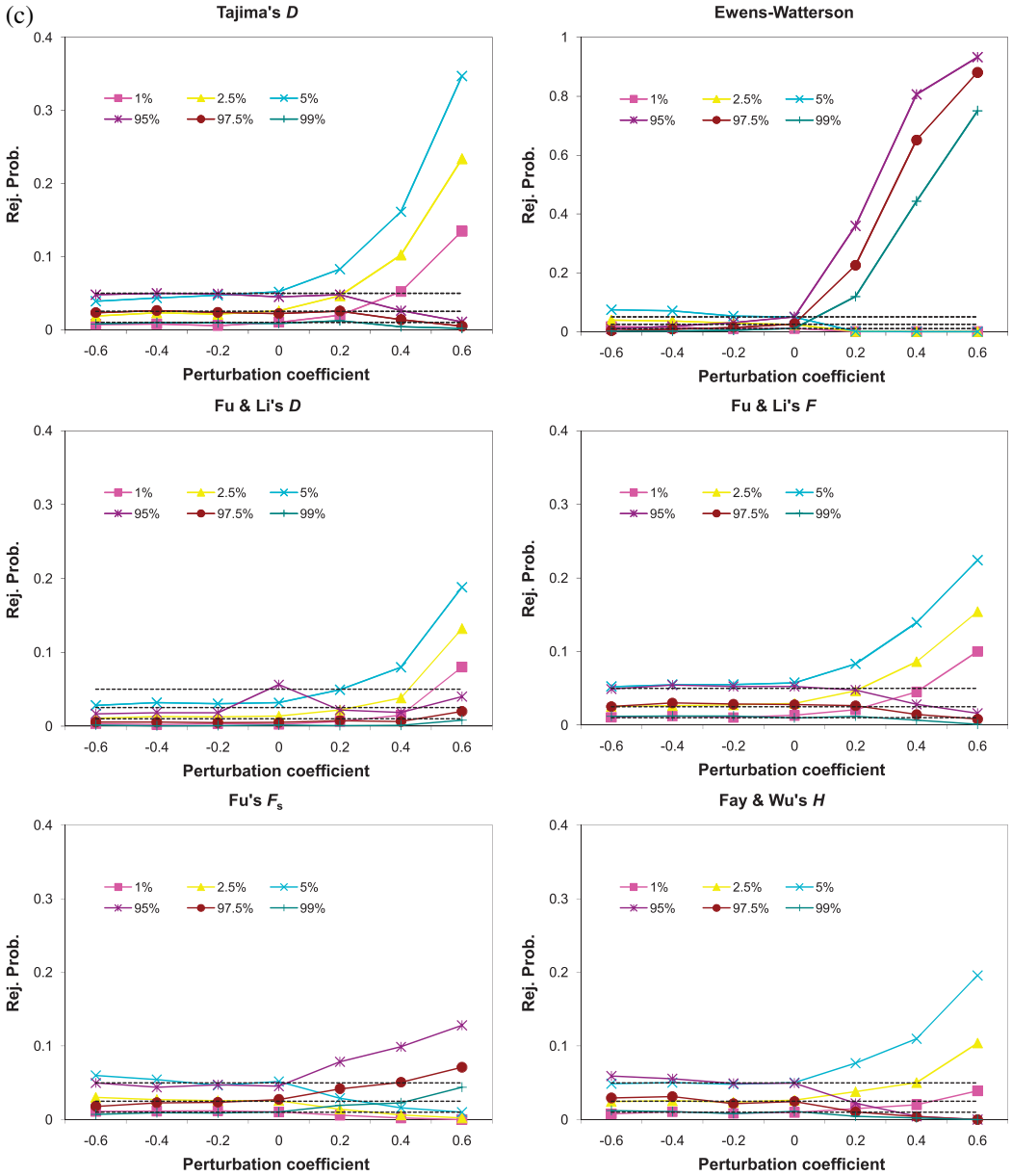

Fig. 2. (Continued).

Usually, it is impossible to determine in nature, for any particular species, whether there are significant cyclical changes in effective breeding size and effective migration patterns, and for which long-term genetic information exists. There is, however, a body of data for which genetically relevant information has been gathered and which made possible an inference about demographic change. These data and inferences are contained in the work of Ives (1945, 1954, 1970) and Band and Ives (1961, 1963, 1968) on wild populations of *Drosophila melanogaster* from the vicinity of South Amherst (MA). *Drosophila melanogaster* in this region undergoes a yearly cycle in abundance. In late spring, flies begin to emerge from small overwintering dormant populations, and successive generations appear in increasing abundance until the late summer and fall when the flies become less numerous and, with the advent of winter, do not appear in traps. Flies were collected at widely separated traps, in samples taken over the seasonal change in successive years, and separate lines were established from each male collected. Recessive lethals on chromosome II were extracted from each line and tested for allelism with the lethals from the other lines. The results over many years of such sampling were that in the spring and early summer the lethals extracted from a given trap showed high allelism with each other, indicating an origin from a recent common ancestor. In the fall collections, however, the allelism of lethals dropped dramatically indicating that the flies from an individual trap now came from a wide variety of ancestors. This pattern was repeated in the next year, but in addition, it was found that there was very low allelism of lethals between years. From these data, Ives and Band (1986) concluded that each year, at end of the breeding season, there was a severe reduction in local population size, with different second chromosomes being

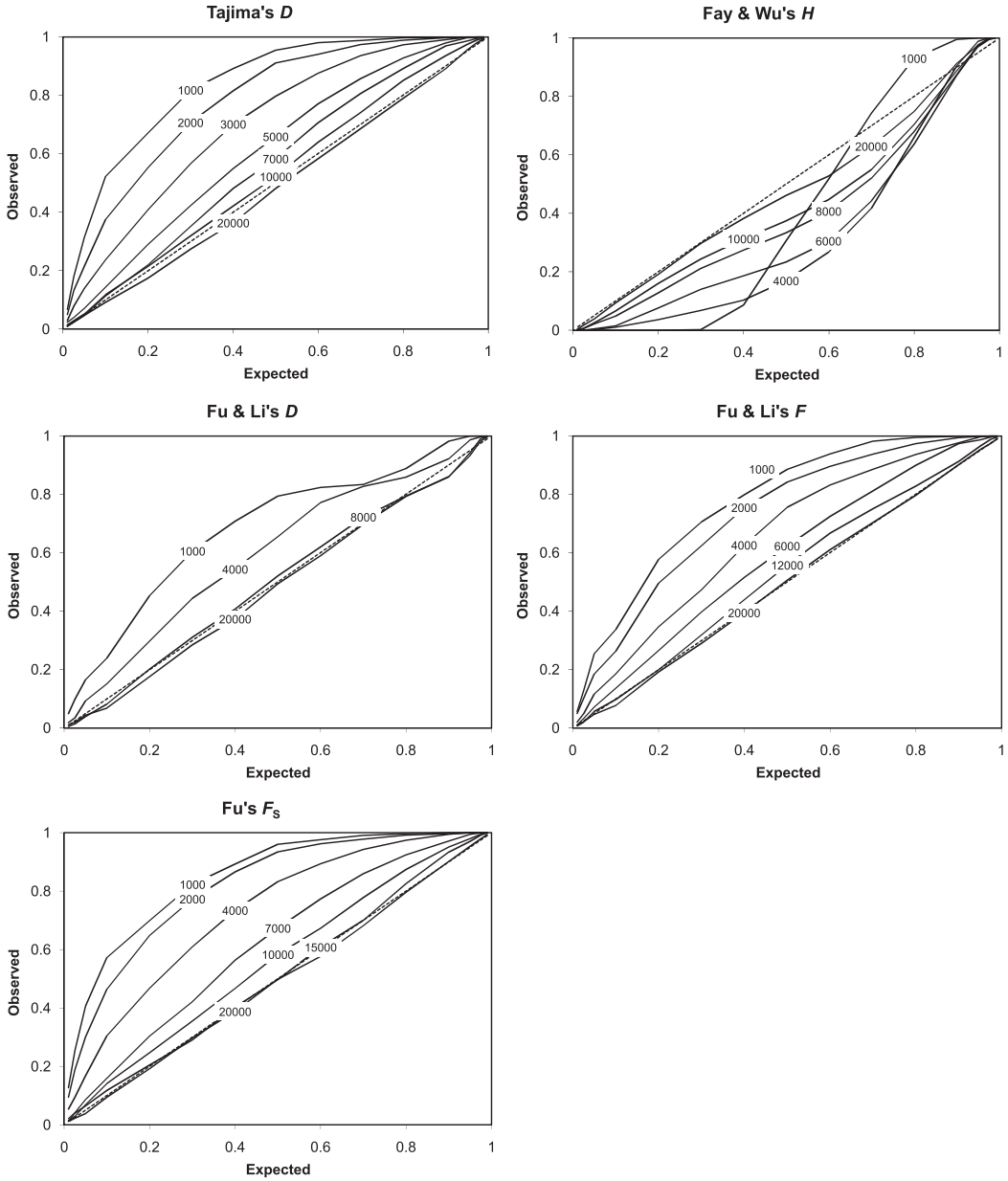

Fig. 3. Comparison of observed and expected cumulative probability distributions of five tests for different numbers of successive generations of relaxation after starting at monomorphism. Numbers on the curves are number of generations; dotted line marks equality of expected and observed. The sample size is $n = 50$, the population size is $2N = 5{,}000$, and the population mutation rate is $\theta = 2$.

preserved at random in each small local relic population. In the spring and summer, these small local populations would then increase in size over the generations of the favorable climate, and mating would occur between individuals derived from different local overwintering ancestors. We then have a cycle, repeated every year, of the establishment in the fall of multiple small overwintering populations sampled from a large interbreeding population that has been produced in the spring and summer from migration among previous small overwintering local populations. This so-called "Ives scenario" may be typical of many species in seasonal climates.

We have modeled examples of this scenario to explore the effect of performing various tests for selection on a species, like *D. melanogaster*, living in an annually cycling environment. A full exploration of even the simplest Ives scenario would involve a range of values for the number of local breeding sites making up a metapopulation, the overwintering size of these populations, the number of such populations, the growth rate of local population size during the breeding season, and a pattern of migration during the year. For the examples given here, each annual cycle consists of 10 successive generations beginning with 10 isolated subpopulations, each of overwintering size 50. There are 10 successive generations within an annual cycle before the winter crash. Each local population grows at a multiplication rate of 1.35, each generation with a mutation rate of 2×10^{-4}, and receives migrants from the metapopulation as a whole at a rate m. Every generation, a sample is taken from each local population and from the metapopulation

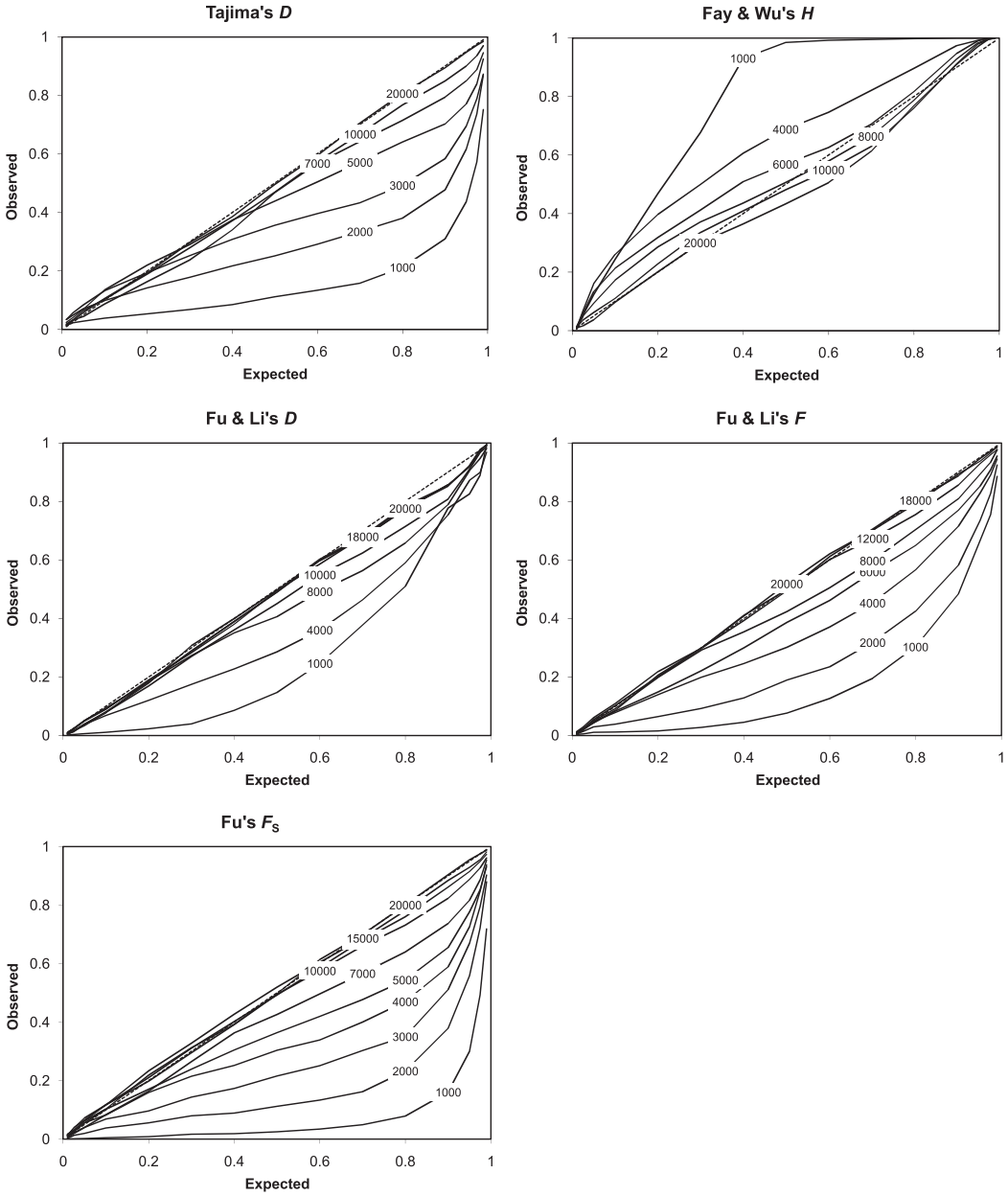

Fig. 4. Comparison of observed and expected cumulative probability distributions of five tests for different numbers of successive generations of relaxation after starting at two equally frequent alleles. Numbers on the curves are number of generations; dotted line marks equality of expected and observed. The sample size is $n = 50$, the population size is $2N = 5{,}000$, and the population mutation rate is $\theta = 2$.

as a whole for the statistical tests. The results of examples with $m = \{0.01, 0.02, 0.05\}$ are given in table 5 after 100 and 1,000 annual cycles (1,000 and 10,000 breeding generations) from the colonization of the area by an originally homozygous migrant pool.

The most consistently unreliable test is the Ewens–Watterson multiple alleles test with excesses of between two and nine times as many significant tests at the lower end of the test distribution at the nominal levels (table 5); the excess showing a tendency to increase slightly with decreasing migration rate. When $m = 0.05$, the tests give roughly one and one-half to two times as many significant results as the nominal values at the lower end of the sample distribution. Fay and Wu's H test is the most conservative at the higher migration rates. At very low migration rates (0.01 and, not shown, 0.001 and 0), all the sites tests are useless in this model because the individual local populations are usually homozygous as a consequence of their small population size over a number of generations and the lack of introduction of variation from migration.

Discussion

The aim of the study was to determine the sensitivity of various statistical tests for selection to perturbations from the theoretical neutral steady state, without specifying the cause of such perturbations. The various perturbations of population composition from the theoretical steady-state neutral distributions show sensitivity of tests of selection to

Table 3. Upper and Lower Tails of the Observed Test Statistic Distributions during Relaxation from an Initially Monomorphic Population of Size $2N = 5{,}000$ from a Sample of Size $n = 50$ and Mutation Rate of $\theta = 2$.

		Lower			Upper		
Test	Generation	0.01	0.025	0.05	0.01	0.025	0.05
Tajima's *D*	1,000	0	0	0.001	0.314	0.180	0.067
	2,000	0	0	0	0.216	0.132	0.050
	3,000	0	0	0.001	0.138	0.078	0.029
	4,000	0	0.002	0.003	0.111	0.059	0.019
	5,000	0	0.002	0.003	0.071	0.040	0.023
Fu and Li's *D*	1,000	0	0	0	0.166	0.099	0.050
	2,000	0	0	0.005	0.146	0.075	0.049
	3,000	0	0	0.011	0.096	0.043	0.027
	4,000	0	0.003	0.013	0.093	0.035	0.018
	5,000	0	0.005	0.030	0.075	0.029	0.013
Fu and Li's *F*	1,000	0	0	0	0.254	0.128	0.057
	2,000	0	0	0.001	0.185	0.102	0.050
	3,000	0	0	0.001	0.122	0.062	0.028
	4,000	0	0	0.003	0.116	0.049	0.020
	5,000	0	0.003	0.012	0.083	0.039	0.016
Fay and Wu's *H*	1,000	0	0	0	0	0	0
	2,000	0	0	0.012	0	0	0
	3,000	0	0.002	0.010	0	0	0
	4,000	0	0	0.011	0.002	0	0
	5,000	0	0.003	0.019	0.004	0	0
Fu's F_S	1,000	0	0	0	0.406	0.255	0.127
	2,000	0	0	0.005	0.300	0.189	0.095
	3,000	0	0	0	0.201	0.125	0.066
	4,000	0	0	0	0.168	0.095	0.054
	5,000	0	0.002	0.013	0.121	0.068	0.031
Ewens–Watterson *F*	1,000	0	0	0	0.174	0.108	0.035
	2,000	0.003	0.009	0.019	0.118	0.066	0.028
	3,000	0.012	0.028	0.045	0.072	0.045	0.019
	4,000	0.004	0.013	0.034	0.086	0.051	0.024
	5,000	0.011	0.027	0.049	0.056	0.034	0.007

deviations of the distribution of allele frequencies in two ways. First, deviations from the theoretical neutral steady state, such that the allele frequency distribution remains L-shaped, are detectable by the tests examined. Deviations in which the most frequent allele becomes yet more frequent, exaggerating the inequality of frequencies (positive perturbations), are more likely to cause statistically significant deviations than are perturbations that move the distribution of genotype frequencies toward more equality (negative deviations). At a low value of θ, however, both positive and negative deviations are detectable. Fu's F_S is generally the least sensitive test over the entire range of deviations; this result contrasts with those of Ramos-Onsins and Rozas (2002) and Ramírez-Soriano et al. (2008), both of whom found Fu's F_S to be the most powerful of the infinite-sites tests, even in the absence of recombination. Both positive and negative deviations may, indeed, be the result of selective events, but they will also arise from a variety of neutral demographic scenarios. Selection in favor of one allele will exaggerate the L shape but so will a sampling event resulting from reduction in population size or gene flow into the population of invaders from a more homozygous donor population, which itself may owe its more extremely asymmetrical distribution to other past historical selective or nonselective causes. On the other hand, an excessive evenness of the allele frequency distribution may arise from the introduction by wholesale immigration of a previously absent allele from a divergent population with a very high frequency of that allele from whatever cause. The massive immigrations and gene exchanges that have characterized human populations during the last 100 generations may certainly give rise to such cases.

Perhaps more important than the sensitivity of tests to the exact shape of the asymmetrical genotypic distribution is the very long relaxation time from an original perturbation to the allele frequency distribution that gives test results that do not yield a significant excess of test observations in the tails of the test statistic distribution. We examined the relaxation process from a diversity of initial frequency distributions that were intended to represent convergence toward the steady state from divergent parts of the frequency state space. The relaxation process characteristically requires on the order of $4N$ generations, irrespective of the initial perturbed state. This insensitivity to the initial perturbed state reflects the fact that a L-shaped distribution of frequencies is achieved well before $4N$ generations by recurrent mutation and random gamete sampling, whether one begins at complete homozygosity or with 50 equally frequent alleles. For a completely homozygous initial state, the process is limited by the accumulation of mutations, but by the time $2N$ generations have passed, there have been $4N^2\mu = N\theta$ mutations in a Poisson-like sampling process from generation to generation. In the case of an initial state of 50 equally frequent alleles, the

Table 4. Upper and Lower Tails of the Observed Test Statistic Distributions during Relaxation from an Initial Condition of two Equally Frequent Alleles in a Population of Size $2N = 5{,}000$ from a Sample of Size $n = 50$ and Mutation Rate of $\theta = 2$.

		Lower			Upper		
Test	Generation	0.01	0.025	0.05	0.01	0.025	0.05
Tajima's *D*	1,000	0.248	0.427	0.563	0.029	0.023	0.014
	2,000	0.130	0.263	0.384	0.071	0.049	0.034
	4,000	0.078	0.159	0.138	0.075	0.056	0.031
	7,000	0.054	0.111	0.150	0.044	0.032	0.016
	10,000	0.031	0.064	0.099	0.064	0.033	0.018
Fu and Li's *D*	1,000	0.006	0.111	0.175	0.006	0.003	0.001
	2,000	0.011	0.223	0.150	0.013	0.006	0.003
	4,000	0.032	0.101	0.128	0.037	0.015	0.005
	7,000	0.027	0.062	0.103	0.028	0.007	0.003
	10,000	0.015	0.033	0.090	0.034	0.016	0.007
Fu and Li's *F*	1,000	0.114	0.245	0.334	0.010	0.005	0.003
	2,000	0.075	0.172	0.265	0.029	0.013	0.005
	4,000	0.055	0.110	0.172	0.048	0.023	0.007
	7,000	0.048	0.084	0.133	0.037	0.015	0.003
	10,000	0.026	0.061	0.100	0.048	0.022	0.007
Fay and Wu's *H*	1,000	0	0	0	0	0	0
	2,000	0	0	0	0	0	0
	4,000	0	0.001	0.007	0	0	0
	7,000	0	0.002	0.023	0.002	0	0
	10,000	0	0.005	0.020	0.004	0	0
Fu's F_S	1,000	0	0	0	0.123	0.054	0.014
	2,000	0	0	0.005	0.185	0.087	0.016
	4,000	0	0	0	0.160	0.069	0.009
	7,000	0	0	0	0.094	0.037	0.009
	10,000	0	0.002	0.013	0.062	0.022	0.012
Ewens–Watterson *F*	1,000	0.011	0.016	0.037	0.008	0.006	0.003
	2,000	0.005	0.012	0.032	0.078	0.049	0.019
	4,000	0.016	0.028	0.045	0.095	0.063	0.017
	7,000	0.012	0.025	0.041	0.073	0.040	0.011
	10,000	0.010	0.023	0.049	0.095	0.058	0.016

sampling process gives rise to a L-shaped distribution almost immediately. The remaining generations of the $4N$ generation relaxation process are then insensitive to the original state.

We must then relate these relaxation times to a realistic view of the frequency of significant demographic perturbations in real populations, in particular, those populations that have been of interest in evolutionary genetics, if the various statistical tests are to be taken simply as tests of selection. In the case of *Homo sapiens*, for example, there is an extreme deviation from the assumption of demographic steady state. Only approximately 100 generations have passed since the Roman Republic, and during that period, there have been repeated nonequilibrium events, including conquests, mass migrations, admixture between geographically disparate populations, expansions, and contractions of local populations. Therefore, it is unlikely that the statistical tests for selection, of the kind considered here, would be informative with respect to the occurrence of natural selection in *H. sapiens*.

Various Drosophila species, especially *D. melanogaster* (but also other widespread Drosophilids that occupy regions with extreme seasonally fluctuating temperature regimes like North America and Europe and which are in part or wholly commensal with humans) are demographically very unstable. *Drosophila melanogaster* is an African species introduced into the Western Hemisphere during the slave trade and repeatedly reintroduced by transatlantic commerce. In the role of commensal, the species has spread throughout the hemisphere via human-assisted migration and is regularly carried from locality to locality. Superimposed on migration is the Ives scenario of annual severe contraction of local population sizes over the winter followed by multiple generations of population growth and movement between local demes during the favorable growth season. As shown in the previous section, if migration rates are low, tests for selection performed on samples taken from a local population deviate significantly from the expected null distribution. Nor should it be assumed that "wild" species of Drosophila do not have some populations with similar demography. *Drosophila pseudoobscura* and *Drosophila persimilis* have been reared out of piles of rotting orchard fruits in successive years, along with *Drosophila immigrans*, *Drosophila hydei*, and *Drosophila busckii*, species that are normally thought of as commensals.

The tests examined in this paper are examples of a class of tests that depend on deviations of the distribution of allelic frequencies from the theoretical steady-state frequency distribution. Another class of tests, exemplified by the MK test (McDonald and Kreitman 1991), the HKA test (Hudson et al. 1987) and d_N/d_S tests (Hughes and Nei 1988) look for evidence of natural selection from comparative differentiation within and between species for both noncoding and coding nucleotide substitutions. At

Table 5. Upper and Lower Tails of the Observed Test Statistic Distributions for the Ives Scenario after 1,000 Cycles (1 cycle is equivalent to 10 generations).

			Lower			Upper		
Test	*m*	Generation	0.01	0.025	0.05	0.01	0.025	0.05
Tajima's *D*	0.01	1,000	0.229	0.234	0.246	0	0	0
		10,000	0.361	0.387	0.420	0	0	0
	0.02	1,000	0.006	0.029	0.124	0	0	0
		10,000	0.003	0.020	0.070	0	0	0
	0.05	1,000	0.005	0.027	0.137	0	0	0
		10,000	0.009	0.024	0.086	0.014	0.003	0.001
Fu and Li's *D*	0.01	1,000	0.029	0.035	0.036	0.006	0	0
		10,000	0.010	0.011	0.012	0.384	0.384	0.163
	0.02	1,000	0.047	0.129	0.129	0	0	0
		10,000	0.023	0.065	0.067	0.001	0.001	0
	0.05	1,000	0.047	0.131	0.133	0	0	0
		10,000	0.028	0.073	0.075	0.001	0	0
Fu and Li's *F*	0.01	1,000	0.030	0.035	0.038	0.079	0.039	0.018
		10,000	0.009	0.011	0.017	0.362	0.362	0.221
	0.02	1,000	0.039	0.127	0.131	0	0	0
		10,000	0.018	0.064	0.070	0.011	0.005	0.001
	0.05	1,000	0.031	0.129	0.138	0	0	0
		10,000	0.020	0.072	0.078	0.014	0.003	0.001
Fay and Wu's *H*	0.01	1,000	0.008	0.017	0.027	0.019	0.004	0.001
		10,000	0.227	0.334	0.395	0.026	0.026	0.004
Fay and Wu's *H*	0.02	1,000	0.001	0.003	0.004	0.002	0	0
		10,000	0.007	0.029	0.027	0.005	0	0
	0.05	1,000	0	0.001	0.007	0	0	0
		10,000	0.009	0.032	0.069	0.004	0	0
Ewens–Watterson *F*	0.01	1,000	0.123	0.129	0.136	0.039	0.028	0.011
		10,000	0.092	0.096	0.102	0.063	0.029	0.011
	0.02	1,000	0.105	0.108	0.120	0.022	0.012	0.002
		10,000	0.071	0.075	0.082	0.064	0.046	0.017
	0.05	1,000	0.096	0.103	0.113	0.015	0.008	0.004
		10,000	0.081	0.088	0.098	0.051	0.031	0.013

NOTE.—The Ives scenario is an example of a regularly-cycling demographic regime. The overwintering population size is $N = 50$, and the population grows to size $N = 1{,}000$. There are 10 populations in total and symmetrical migration occurs with rate m in each generation. The mutation rate per generation per locus is $\mu = 0.002$. The sample consists of 50 chromosomes pooled from samples of 5 from each local population.

first sight, these latter tests should not be sensitive to assumptions about demography, but Eyre-Walker (2002) has claimed that the MK test is sensitive to changes in the effective population size, and Ingvarsson (2004) has shown that the HKA test can yield significant results as a result of population subdivision. Furthermore, the power of tests that rely exclusively on the relative frequencies of replacement versus silent nucleotide substitutions depends upon the consistency with which natural selection acts over timescales that generally exceed the life span of an individual species (Garrigan and Hedrick 2003).

The basis for a third form of test for natural selection is the simple "hitchhiking" of Maynard Smith and Haigh (1974). This form is generally used to identify high-frequency haplotypes that exhibit long-range linkage disequilibrium, a pattern that suggests a rapid increase in frequency of a new mutation. This method has been applied to human population data (Sabeti et al. 2002), and the "extended haplotype homozygosity" test compares the observed haplotype frequencies and levels of linkage disequilibrium with distributions generated by Monte Carlo simulation. A simulation study making explicit use of a best-fit "out-of Africa" demographic scenario shows reasonably close correspondence between a nonselective scenario and the observed distribution of allele frequencies, linkage disequilibrium, and population differentiation (Schaffner et al. 2005). There is a large variety of specific historical demographic scenarios for various human groups at different times in the past. Other investigators have demonstrated that alternatives to the simple hitchhiking model of Maynard Smith and Haigh (1974), such as selection on standing neutral variation, result in diminished levels of power for this and other categories of test (e.g., Przeworski et al. 2005). The question still to be investigated is whether the variety of tightly linked sequences of nucleotides can be rigorously distinguished as a consequence of selection.

Conclusions

We are left, then, with the problem of making inferences about selection from statistical tests on standing genetic variation. Ordinarily, we require that robust tests of selection must have markedly greater power to detect selectively caused deviations from the null distribution than other, nonselective, causes. However, the demonstration of the relative power of tests to detect various specific alternative departures from the null structure can only be determined by the examination of the outcome of various detailed

parametric models. To measure the relative power of a test to detect selection as opposed to, say, regular expansions and contractions of local populations, it is necessary to have both exact parameterized models of the kinds of selection to be detected and of models of demographic change. To decide between the alternative explanations would then require actual estimates in nature of the demographic variables, an almost impossible task for most species in nature.

A three-step alternative is suggested by considering the investigations of Schaffner et al. (2005) and by Sabeti et al. (2002) discussed above. In the first step, a large number of genomic regions is chosen for sequencing in the sample of organisms from nature, and test statistics are calculated for each region, yielding a distribution of empirical values of the test statistics across genomic regions that does not depend on any a priori null structure but, instead, reflects the actual demographic and historical effects that are not region specific. In the second step, extreme outliers in this empirical distribution are chosen as candidates for locus-specific effects, attributable to either selection on the sequence itself or selection at closely linked sites. In the third step, new samples are taken to confirm the status of the extreme outlying cases. Although the original work involved in sequencing many regions is considerable, the yield of potential sites of selection is not confined to a single a priori case. What such a procedure provides, however, is not an identification of loci under selection but an enriched list of candidate loci. Moreover, as shown by Przeworski et al. (2005), some forms of selection have no net effect on the final distribution of allelic frequencies, so these would be missed by the screen. Loci that do fall in the tails of these empirical genomic distributions may, at best, represent the subset of non-neutrally evolving loci that adhere to the assumptions of a simple hitchhiking model. Thornton and Jenson (2007) offer several suggestions for how the false-positive rate discovery rate may be reduced in genomic scans for natural selection. Yet, in general, an unambiguous demonstration of selection requires allele-specific physiological and fitness measurements. But this raises a new set of problems. First, such biological investigations could only demonstrate either selection still in progress or, less rigorously, a likely selective difference in response to conditions known to have occurred in the past, as, for example, a major past lethal disease epidemic. Second, there are very few species in which age-specific mortality and reproductive data under natural conditions are available. From this standpoint, humans would be the most satisfactory, but this is precisely the species for which an assumption of demographic stability over a long period may be least applicable.

Acknowledgments

The work was supported by National Institutes of Health (NIH) grant (NIH-GM070543 to R.L. and J.W.).

References

Band HT, Ives PT. 1961. Correlated changes in environment and lethal frequency in a natural population of *Drosophila melanogaster*. *Proc Natl Acad Sci USA*. 47:180–185.

Band HT, Ives PT. 1963. Comparison of lethal + semilethal frequencies in second and third chromosomes from a natural population of *Drosophila melanogaster*. *Can J Genet Cytol*. 5:351–357.

Band HT, Ives PT. 1968. Genetic structure of populations IV. Summer environmental variables and lethal and semi-lethal frequencies in a natural population of *Drosophila melanogaster*. *Evolution*. 22:633–641.

Christiansen FB, Frydenberg O. 1973. Selection component analysis of natural polymorphisms using population samples including mother-child combinations. *Theor Popul Biol*. 4:425–445.

Ewens WJ. 1969. Population genetics. Methuen, London.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol*. 3:87–112.

Ewens WJ, Gillespie JH. 1974. Some simulation results for the neutral allele model, with interpretations. *Theor Popul Biol*. 6:35–57.

Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics*. 162:2017–2024.

Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics*. 155:1405–1413.

Fu Y-X, Li W-H. 1993. Statistical tests of neutrality of mutations. *Genetics*. 133:693–709.

Fu Y-X. 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics*. 143:557–577.

Fu Y-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*. 147:915–925.

Garrigan D, Hedrick PW. 2003. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution*. 57:1707–1722.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics*. 116:153–159.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*. 335:167–170.

Ingvarsson PK. 2004. Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Genet Res*. 83:31–39.

Ives PT. 1945. Genetic structure of American populations of *Drosophila melanogaster*. *Genetics*. 30:167–196.

Ives PT. 1954. Genetic changes in American populations of *Drosophila melanogaster*. *Proc Natl Acad Sci USA*. 40:87–92.

Ives PT. 1970. Further genetic studies of the South Amherst population of *Drosophila melanogaster*. *Evolution*. 24: 507–518.

Ives PT, Band HT. 1986. Continuing studies on the South Amherst *Drosophila melanogaster* natural population during the 1970's and 1980's. *Evolution*. 40:1289–1302.

Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*. 176:2371–2379.

Kim Y. 2006. Allele frequency distribution under recurrent selective sweeps. *Genetics*. 172:1967–1978.

Kimura M, Crow JF. 1964. The number of alleles that can maintained in a finite population. *Genetics*. 49:725–738.

Kreitman M. 2000. Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet*. 1:539–559.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res*. 23:23–35.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*. 351:652–654.

Nielsen R. 2001. Statistical test of selective neutrality in the age of genomics. *Heredity*. 86:641–647.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet*. 39:197–218.

Orr HA. 2009. Fitness and its role in evolutionary genetics. *Nat Rev Genet*. 10:531–539.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution*. 59:2312–2323.

Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*. 179:555–567.

Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. *Mol Biol Evol*. 19:2092–2100.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 419:832–837.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics*. 132:1161–1176.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*. 15:1576–1583.

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*. 141:413–429.

Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M. 2005. On the meaning and existence of an effective population size. *Genetics*. 169:1061–1070.

Tajima F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics*. 123:597–601.

Tajima F. 1989b. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 123:585–595.

Thornton K. 2005. Recombination and the properties of Tajima's *D* in the context of approximate-likelihood calculation. *Genetics*. 171:2143–2148.

Thornton KR, Jenson JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics*. 175: 737–750.

Watterson GA. 1978. The homozygosity test of neutrality. *Genetics*. 88:405–417.

Zeng K, Shi S, Wu C-I. 2007. Comparisons of site- and haplotype-frequency methods for detecting positive selection. *Mol Biol Evol*. 24:1562–1574.